SYNOPSIS OF THE THESIS ENTITLED

"IMPROVING THE LEVEL OF PROTECTION IN IOT ENVIRONMENTS BASED ON MACHINE LEARNING "

SUBMITTED FOR THE DEGREE OF DOCTOR OF COPUTER SCIENCE

Under the Faculty of Science & technology

By Research Student

Kahlan Hezam Sarhan Alnafesh

Under the Guidance of

Dr. Sudhir Baburao Jagtap

Principal

Swami Vivekanand

Mahavidyalaya, Udgir

SWAMI RAMANAND TEERTH MARATHWADA
UNIVERSITY, NANDED

2022-2023

1. Abstract

The Internet of Things includes everything from cameras to wearables to Smart factory equipment to household appliances and most of the things in our daily life because of its ability to transmit data over a network to the virtual world. The Internet of Things infrastructure is particularly vulnerable to attack because of its small size, limited storage, hardware, and battery life. Because of these flaws, the Internet of Things is particularly vulnerable to attacks, including botnets, distributed denial-of-service attacks, eavesdropping, Man in the Middle attacks. With the help of a recent IoT dataset, this study aims to examine how well machine learning and deep learning models detect internet of things malware. Machine learning algorithms like XGBoost, Catboost, Random Forest, and Deep learning techniques like a Convolutional Neural Network and Long short-term memory (LSTM) will be tested on the IoT dataset. This study includes four phases: feature extraction, feature selection, classification model, and suggest the effective way to apply the model to the IoT framework.

2. Introduction

Internet of Things encompasses a wide range of devices, from cameras, wearables, Smart factory equipment, household appliances, medical gadgets, and even nuclear reactors [1], because of its ability to transmit data across a network to the virtual world. Small devices with limited storage, hardware, and battery life, as well as the inability to implement security procedures, make the Internet of Things' infrastructure particularly vulnerable. As a result of these vulnerabilities, the Internet of Things is particularly susceptible to attack, including from botnets, distributed denial of service attacks, eavesdropping, Man in the Middle attacks, firmware hijacking. Etc

Any intrusive software developed by hackers to steal data or damage computer systems and physical devices is known as malware [2], short for malicious software." Viruses, worms, spyware, ransomware, Trojan horses and adware" are more common types of malwares. In the wake of recent malware attacks, large amounts of information have been stolen.

Personal computers running on Windows and Linux, the world's most popular operating systems, have been the primary target of malware for the past few decades. Internet of Things (IoT) technology has led to a dramatic increase in the variety of computing devices [3]. Many microcontrollers or CPUs are used in IoT devices, including those running on resource constrained systems. It's because IoT devices lack security design or

implementation. As we mentioned above are becoming a favourite target for hackers. DDoS attacks, port scanning, and brute-force attacks are just some of the characteristics of IoT malware that can be used to gain access to IoT devices.

As a result, researchers are under increasing pressure to strengthen IoT device security, particularly in light of the prevalence of IoT malware. There are a lot of studies on the safety of IoT devices out there. Focused on existing protocols, mechanisms, and traditional techniques like; digital signature encryption to secure IoT communications [4]. In addition, some studies proposed emerging methods like software-defined Networking and blockchain [5] to eliminate the security issues and malware in the IoT platform.

In recent years, as the number of internet-connected devices (IoT) has exploded, so too have the types of attacks that target them. Malware embedded in IoT devices can be complicated to detect, making it one of the most dangerous attacks. Machine learning and deep learning are two of the most effective techniques applied recently for detecting various forms of attacks on the Internet of Things. Because signatures fail to detect unknown malware, machine learning has replaced more traditional approaches based on signatures in most security solutions today.

There are two primary classifications of IoT malware behavior detection strategies: dynamic and static analysis [6]. The dynamic approach entails keeping an eye on executables while running and looking for unusual behavior. It takes a lot of resources to keep track of running processes, and malware can infect natural environments. In addition, because many types of malw require trigger conditions to carry out their malicious actions, it is impossible to fully monitor all of their behaviors during execution. IoT executable files face a wide range of challenges when being run, not the least of which is the wide variety of architectures they can be run on, like MISP ARM. As effective as dynamic analysis is for Windows malware analysis, it has difficulty performing consistent analysis on IoT devices architecture because it executes binary programs in sandbox environments to monitor run-time behavior. The lack of support for on-device dynamic research by IoT devices with limited system resources also makes malware protection based on dynamic analysis infeasible.

Instead of relying on obfuscation and evasive programming techniques[7], most current IoT programs adhere to simple programming principles, unlike recent malware or applications for Windows and Android. Consequently, we can utilize their simplicity to develop a detection and classification method for IoT malware that is effective and efficient.

There has been a significant increase in the development of machine intelligence in the last few years, and it has moved from laboratory curiosity to real-world equipment and application. There are numerous valuable applications. Internet of Things monitoring solves new problems or hasn't even been invented attacks yet. ML And DL are potent methods for data exploration. Finding out what is considered "normal" and "abnormal" behaviour, according to how IoT devices and components work together environment. Therefore, ML/DL methods are critical in this field. Enhancing the security of IoT systems rather than merely facilitating their implementation of security-based communication between two or more devices, so by using machine learning techniques, it is possible to identify new malware with high efficiency and accuracy by analysing the behaviour of the existing malware-based on characterized features or patterns.

When it comes to malware detection, there are three main approaches: signatures, behavior, and specification-based methods. Signature is the files scanned and checked against an existing list using the signature technique; those that match the list are considered malicious. Instead of examining the program's signature, the behavioral-based approach looks at how it behaves.

Traffic traces and network behaviour of IoT devices will be analysed in this study, which uses attributes from traffic patterns based on different protocols to identify malware, then apply the classification techniques on malware datasets to build a robust trained model able to detect different types of attacks. Finally, we aimed to deploy the model in the gateway or another platform like SDN to detect and mitigate the malware in real-time.

3. Related work

[8] Experimentation is used in the design science research method. After deciding on an IoT dataset (Bot-IoT) and nine machine learning classifiers, the implementation moves to step two of three: feature engineering, which involves extracting and encoding features. Optimization of the learning algorithm and improvement of detection through attribute conversion and feature combination execution speed, (c) classification, and comparative analysis for ML evaluation similar works that already exist as models. The obtained result was 99.9 % Accuracy using the LR classifier.

[9] The authors demonstrated how the Internet-of-Things (IoT) is vulnerable to malware attacks because of its implementation and autonomous operation features. Understanding IoT malware behaviour for mitigation and prevention can be accomplished through static analysis of executable files for that malware. The current analytic approaches that use opcodes or call graphs are often resource-intensive and do not work well with many CPU

architectures. Machine-learning techniques have been used in this study to detect and classify IoT malware programs based on Byte Sequences from Executable File. The data set was collected from multi resources like Virus Total and m commercial IoT-device vendors like; Netgear MikroTik. The samples consisted of 111K benign malware and 111K malware. The proposed malware detection reached 99.96% accuracy by SVM and for malware reached 98.47% of accuracy by SVM classifier.

Malware Threat Hunting System proposed by [10] for malware detection in the IoT platform. First," the malware binary is converted into a color image, and then the machine or deep learning analysis is conducted for efficient malware detection. Finally, establish a reference point against which MTHS's performance can be evaluated through traditional methods for detecting and preventing malware". Two Windows and Android software have been able to access public datasets. Compared to the results of the previous study, the experimental results show a positive outcome. The mailing dataset contained 9339 malware samples belonging to 25 families and 617 non-malignant samples. The data was gathered from the University of California's Vision Research Lab. 14,733 malware samples and 2486 benign samples of different Android IoT applications were in a single Leopard Mobile dataset. The study was conducted using a Deep Convolutional Neural Network, and the highest accuracy was 98.18%.

[11] Decentralizing, distributing, and transferring computation to IoT nodes are all goals of edge computing to support the Internet of Things. In addition, IoT devices that Cybercriminals primarily use perform attacks that exploit applications on the Internet of Things (IoT) networks. The authors proposed reliable methods to identify nodes engaging in malicious activity if the network is kept safe. Fuzzy and fast fuzzy pattern tree methods detect malware by converting OpCodes into a vector space—the short and fuzzy pattern tree for extracting and classifying characteristics. Vx-Heaven, a Kaggle dataset, a Ransomware dataset, and an IoT dataset were all utilized. SVM, Random Forest, and Decision Tree were used to achieve a precision of 100 percent on the Vx-Heaven dataset.

"IoT Malware Classification Based on System Calls" was proposed by [12]. The method uses IoT malware behavior classification and clustering via system call monitoring. The malware samples have been called from honeypot and analyzed: Learn the behavior of the new class of malware by detecting and detecting known malware. To discover previously unknown classes, then apply classification and clustering techniques. Three thousand nine hundred eighty-seven malware samples have been used. The following method was: Assign known malware to a family and reject unknown malware using multiple One-class SVM

models. Unknown malware that behaves, in the same way, should be assigned to clusters. Clusters that have a sufficient amount of data will be categorized as a new family of malware. To model a family of unknown malware that the Clustering component has detected, create a one-class SVM model. The metrics were as follows; SVM classifiers for known malware with F1-Score reached 98% and for unknown malware reached 97%.

[13] Has proposed a method for Malware Detection in Fog Layer. This method has three important phases for pre-processing: feature extraction, feature selection to reduce the number of features, and classification. Convert the binary files to hexadecimal using HexDump, segmentation into 4-gram, features selection and reduction using Gain Ratio; decision tree and Component Analysis (PCA), Finally apply decision tree classifier. The accuracy was Acc. 96.7.

[14] Have presented a new method based on machine learning for detecting Mirai malware "NBaIoT" dataset, which data consist of features infected by the Mirai Malware, is used in that study. The Cross-Validation technique has been used for data splitting to overcome overfitting, and the experiment was conducted using ANN. The achieved accuracy is 92.8%

[15] The Opcode dataset has been used in this research; it consists of 70,140 normal and 69,860 malicious malwares. The IoT Device dataset's benign or malignant input is classified using a deep neural network (DNN). The obtained accuracy reached 99.7 %

4. Problem statement

Malware is malicious code designed with the intent to gain access and damage your computer or IoT device. With the spread of the Internet of Things devices, experts have found that IoT technology can be hacked by malware. These devices are always connected to the internet and don't have a lot of safety. Because IoT devices are small and lightweight, they don't have the power to run traditional antimalware programs. Hackers can easily access them because of this. Hackers are increasingly turning to malware attacks as their primary method of attack.

In some cases, they have illegally used IoT malware attacks to mine cryptocurrency. One of the most famous attacks that control malware has spread widely in previous years, like Botnet that exploited hundreds of thousands of IoT devices. Like It's not malware, but infecting devices is how this attack is carried out. It's a network of robots. Hackers use malware to infiltrate an existing network and turn it into a botnet. These devices are under

the control of hackers at all times. In the case of a Mirai attack, Reaper, for example, a virus spreads itself.

Malware that infects IoT devices by using factory default login information infects unprotected devices such as IoT.

Malware detection and prevention using machine learning and deep learning techniques are standard. However, designing an effective ML model that accurately detects and prevents IoT malware is tricky. In addition, the lack of new IoT datasets complicates the design and development of systems to detect malware in IoT networks. To determine the effectiveness of malware intrusion detection systems, it is critical to evaluate the few available datasets.

Malware is a source of attacks on the system's confidentiality, integrity, authenticity, and availability. The current work aims to investigate the effective method to detect malware in IoT with Machine learning and Deep learning. **To achieve this, we have stated our objectives like the following:**

1. To identify the IoT devices by analysis of their operating patterns using different intelligent learning approaches.

2. To characterize malware behaviors based on the analysis of IoT network traffic with a high level of accuracy.

3. To design and implement a reliable and cost-effective malware detection system in the IoT environments based on feature engineering of network traffic.

5. Methodology

When applying security solutions to IoT devices, the low power retention and computational processing capabilities of these devices pose a significant challenge. This is why traditional antimalware systems like; Static analysis doesn't work well in the Internet of Things. This study aims to determine whether or not a new unknown sample is malicious and then classify it according to its characteristics from among the various families of malware that are already known. To achieve our goals, we will follow this procedure.

- Conducting a comprehensive survey of available studies that use traditional or recent techniques to detect the malware and prevent it.

To achieve our goals, we will follow this procedure.

5.1 Data description

After searching the web and other sources, we found that the data that would be appropriate and achieve the purpose of the study was to detect the malware in the Internet of Things. We found the IoT-23 dataset was the best choice.

Dataset IoT-23 produced by Avast AIC laboratory. From various IoT devices, there are 20 malware captures as well as 1 benign abnormality in the collection. IoT-23 is a collection of network traffic from the Internet of Things (IoT) devices. Data collected from IoT malware infections and benign IoT traffic. The dataset consists of 22 features including the labels that are consist of 10 classes.

5.2 Pre-processing

Data pre-processing is the process of transforming unstructured data into a form that can be read, accessed, and analysed. The raw data cannot be used in data mining; therefore, this is a critical step to get acceptable and good performance. Check the data quality before using machine learning or other methods.

The steps will be the following:

1- Make the dataset in a readable format.
2- Remove the unnecessary symbols
3- Fill in the missing value and NAN values.
4- Remove the duplicate values.
5- Scaling the data to be suitable and feedable to model

5.3 Feature Selection

Feature selection has a significant influence on the success of our model. Training machine learning models on data attributes can have a significant impact on their performance. It's the process of selecting the features that have the greatest impact on the prediction variable or output that we are interested in. In our experiment, we will use some techniques like correlation and features selection algorithms to find the important feature that will affect the performance mode

5.4 Data Balancing

Machine learning algorithms may encounter difficulties in situations where the learning dataset has an imbalance in class distribution. In the case of data imbalance, under-sampling

the majority class is a common strategy. A well-balanced dataset can be achieved through the use of over-and under-sampling techniques so we will use some techniques to overcome this issue like SMOTE.

5.5 ML and DL training and test

Both ML and DL use data segmentation to create training, test, and validation sets. We will use two techniques traditional and K-fold.

1- Tradition splitting technique by using 70 % of data as training set and 30 % as testing set.

2- k-Fold Cross-Validation will be used for splitting data to overcome overfitting and underfitting

5.6 Implementation

We will implement machine learning and deep learning algorithms. For Machine learning, we will use the recent Machine learning algorithm XGBoost, Catboost, Random Forest. And for Deep Learning we will implement CNN, LSTM

5.7 Evaluation

The confusion matrix is applied to evaluate the performance of classification models for a given data set. It is applicable when the test data are known for true values. It appears to be similar to the following table.

	Predicted values Positive	Predicted values Negative
Actual value Positive	TP	FP
Actual values Negative	FN	TN

6. Tools

System Specifications

6.1 Hardware Requirement

1. Processor: Core I7 8th-10th
2. RAM: 16-32 GB
3. Operating System: Windows 11 OS
4. GPU: Graphics card NVidia 4-6gb Gtx 1650 or 2060 RTX.
5. Storage: 1 TB SSD HD

6.2 Software

1. Anaconda (Python)
2. google Collaboratory.
3. Packages: like Tensorflow, Keras. Etc

7. Work details and time constraints

Our work will be organized as follow:

Activity	First Month	Second Month	Third Month	Fourth Month
Identification research area				
Write a research proposal				
Literature review				
Framework formulation				
Designing				
Implementation				
Data collection & data analysis				
Testing & evaluation				
Writing: A review of literature				
Writing: Research methodology				
Writing: Findings				
Writing: Discussion				
Writing: Evaluation				
Proofreading & editing				

Table1: showing details of work and time

8. Anticipated Outcomes

1. Detect and mitigate malware attacks in the primitive phase.
2. Build a robust model able to detect a malicious attack in an IoT environment with high accuracy.
3. Proposed intelligent system architecture for IoT with low cost to be integrated with our AI- model to identify cyber-attacks on IoT in real-time.

References:

[1] Qiu, T., Chen, N., Li, K., Atiquzzaman, M., & Zhao, W. (2018). Thirdquarter 2018,"How can heterogeneous Internet of things build our future: a survey,". *IEEE Commun. Surv. Tutorials*, *20*(3), 2011-2027.E. M. Rudd, A. Rozsa, M. Günther, and T. E. Boult, "A Survey of Stealth Malware Attacks,

Mitigation Measures, and Steps Toward Autonomous Open World Solutions," *IEEE*

Communications Surveys and Tutorials, vol. 19, no. 2, pp. 1145–1172, Apr. 2017, doi:

10.1109/COMST.2016.2636078.

[2] J. Su, V. Danilo Vasconcellos, S. Prasad, S. Daniele, Y. Feng, and K. Sakurai, "Lightweight Classification of IoT Malware Based on Image Recognition," *Proceedings - International Computer Software and Applications Conference*, vol. 2, pp. 664–669, Jun. 2018, doi: 10.1109/COMPSAC.2018.10315.

[3] M. Alhanahnah, Q. Lin, Q. Yan, N. Zhang, and Z. Chen, "Efficient signature generation for classifying cross-architecture IoT malware," *2018 IEEE Conference on Communications and Network Security, CNS 2018*, Aug. 2018, doi: 10.1109/CNS.2018.8433203.

[4] M. Pourvahab and G. Ekbatanifard, "An efficient forensics architecture in software-defined networking-IoT using blockchain technology," *IEEE Access*, vol. 7, pp. 99573–99588, 2019, doi: 10.1109/ACCESS.2019.2930345.

[5] A. Souri and R. Hosseini, "A state-of-the-art survey of malware detection approaches using data mining techniques," *Human-centric Computing and Information Sciences*, vol. 8, no. 1, pp. 1–22, Dec. 2018, doi: 10.1186/S13673-018-0125-X/FIGURES/8.

[6] S. Hosseinzadeh, S. Hyrynsalmi, and V. Leppänen, "Obfuscation and diversification for securing the internet of things (IoT)," *Internet of Things: Principles and Paradigms*, pp. 259–274, Jan. 2016, doi: 10.1016/B978-0-12-805395-9.00014-9.

[7] W. Yaokumah, J. K. Appati, and D. Kumah, "Machine Learning Methods for Detecting Internet-of-Things (IoT) Malware," *International Journal of Cognitive Informatics and Natural Intelligence*, vol. 15, no. 4, pp. 1–18, Oct. 2021, doi: 10.4018/ijcini.286768.

[8] T.-L. Wan *et al.*, "Efficient Detection and Classification of Internet-of-Things Malware Based on Byte Sequences from Executable Files," *IEEE Open Journal of the Computer Society*, vol. 1, pp. 262–275, Oct. 2020, doi: 10.1109/ojcs.2020.3033974.

[9] H. Naeem, "Detection of Malicious Activities in Internet of Things Environment Based on Binary Visualization and Machine Intelligence," *Wireless Personal Communications*, vol. 108, no. 4, pp. 2609–2629, Oct. 2019, doi: 10.1007/s11277-019-06540-6.

[10] E. M. Dovom, A. Azmoodeh, A. Dehghantanha, D. E. Newton, R. M. Parizi, and H. Karimipour, "Fuzzy pattern tree for edge malware detection and categorization in IoT," *Journal of Systems Architecture*, vol. 97, pp. 1–7, Aug. 2019, doi: 10.1016/j.sysarc.2019.01.017.

[11] D. Kien, H. Dai, T. Nguyen, and D. Loi Vu, "IoT Malware Classification Based on System Calls," 2020.

[12] B. M. Khammas, "The performance of iot malware detection technique using feature selection and feature reduction in fog layer," in *IOP Conference Series: Materials Science and Engineering*, Nov. 2020, vol. 928, no. 2. doi: 10.1088/1757-899X/928/2/022047.

[13] T. G. Palla and S. Tayeb, "Intelligent mirai malware detection for iot nodes," *Electronics (Switzerland)*, vol. 10, no. 11, Jun. 2021, doi: 10.3390/electronics10111241.

[14] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, and S. Venkatraman, "Robust Intelligent Malware Detection Using Deep Learning," IEEE Access, vol. 7, pp. 46717–46738, 2019, doi: 10.1109/ACCESS.2019.2906934.

Research scolar

Kahlan Hezam Alnafesh

Research Guide Kahlan Hezam Alnafesh

Dr. Sudhir Baburao Jagtap

Principal of Swami Vivekanand

Mahavidyalaya, Udgir

S.R.T.M.U. Nanded

Dr.Sudhir B.Jagtap
Research Guide/Supervisor
Research Centre in Computer Science
Swami Vivekanand Mahavidyalaya,
Udgir Dist.Latur (MH)